Abnormal Events Detection Using Deep Networks for Video Surveillance

Binghao Meng[1], Lu Zhang[1], Fan Jin[1], Lu Yang[1], Hong Cheng[1](✉), and Qian Wang[2]

[1] School of Automation Engineering, Center for Robotics,
University of Electronic Science and Technology of China, Chengdu, China
hcheng@uestc.edu.cn

[2] Ricoh Software Research Center of Beijing, Beijing, China

Abstract. In this paper, a novel method is proposed to detect abnormal events. This method is based on spatio-temporal deep networks which can represent sequential video frames. Abnormal events are rare in real world and involve small samples along with large amount of normal video data. It is difficult to apply with deep networks directly which usually require amounts of labeled samples. Our method solves this problem by pre-training the networks on videos which are irrelevant to abnormal events and refining the networks with fine tuning. Furthermore, we employ the patch strategy to improve the performance of our method in complex scenes. The proposed method is tested on real surveillance videos which only contain limited abnormal samples. Experimental results show that the proposed approach can outperform the conventional abnormal event detection algorithm which utilized hand-crafted features.

Keywords: Spatio-temporal networks · Deep learning · Abnormal events detection · Small sample events

1 Introduction

In the past decades, with the development of the technology in computer vision and pattern recognition, smart surveillance systems are increasingly being used to detect potential dangerous situations. But the recognition of complex events in videos continues to be a challenging problem [2,3,9]. For detection of complex events in videos, recent researches in this aspect emphasized on concept based methods as they provided high-level complex event recognition and proposed an approach to discover data-driven concepts to represent high level semantics of video [13,22], so as to improve complex event recognition, [1] developed a model that captured the temporal dynamics of windowed mid-level concept detectors. Based bottom up approach was presented to recognize events [1]. To detect abnormal event in scenes, various approaches were categorized into two classes, one was based on trajectory pedestrians or object-tracking, nevertheless, these methods were not the best choice because tracking was extremely

F. Sun et al. (Eds.): ICCSIP 2016, CCIS 710, pp. 197–204, 2017.
DOI: 10.1007/978-981-10-5230-9_22

difficult in complex scene [18,24]. Another was based on motion representation which avoided tracking, the method of mainstream was optical flow such as [3,7], Markov Random Field (MRF) model such as [14,19]. [12] developed a 3DCNN models, and achieved superior performance in the recognition of human actions.

Deep learning has won amounts of contests in pattern recognition and machine learning, and it allows models to learn representations of data with multiple levels of abstraction. Deep convolution nets have brought about breakthroughs in image and video processing. Recently, deep models were led into video event analysis, as the great performance of Convolution Neural Networks (CNN) in image classification, it was expanded to time domain to temporal clues of videos [12]. Video level pooling or encoding to frame features extracted by CNN were implemented to obtain the video level representation [6]. Another way to extract the video level representation was to combine CNN/3DCNN with Recurrent Neural Networks (RNN) with Long Short-Term Memory (LSTM) [4,20], which has been proved to be efficient in sequential data processing [8]. [5,15] proposed a fully automated deep model which learned to recognize human actions directly from video.

In this paper, we aim to detect abnormal events of crowding and escape behaviors in videos. We are more inclined to post-processing in our emergency plan systems, which lack of beforehand prevention. Mass incident in public place tends to have greater security risks. It is necessary to develop a system which can implement real-time monitoring of the mass incident, judging whether any mass incident, predicting the state of abnormal behavior in a short time, and making corresponding warning or alarm. For abnormal behavior detection, approaches based on Social Force Model (SFM) such as [17,23] has already proved effective. In this work we propose an automated deep learning model that address this problem. The contributions include: (1) solve the problem of small sample abnormal event by pre-training on irrelative videos and fine-tuning on abnormal videos. (2) efficiently detect events which occur in local area of surveillance viewpoint. The rest of this paper is organized as follows: we describe the overview of the approach in Sect. 2. The details of the proposed method described in Sect. 3. The experimental results on the dataset are analyzed in Sect. 4. We conclude in Sect. 5.

2 Overview of the Proposed Framework

We propose a novel abnormal event detection approach using spatio-temporal deep networks. Figure 1 summarizes the main steps of the method. We first extract five low-level features as inputs of different pathway of the deep networks: gray, gradient-x, gradient-y, optflow-x and optflow-y. The gray contains the original information of image for that color information has little impact on event detection. Gradient can represent the edge information efficiently, so the gradient-x and gradient-y are obtained by computing gradients along the horizontal and vertical directions respectively. The optflow-x and optflow-y contain the optical flow fields along the horizontal and vertical directions, respectively, computed from adjacent input frames which contain the motion information.

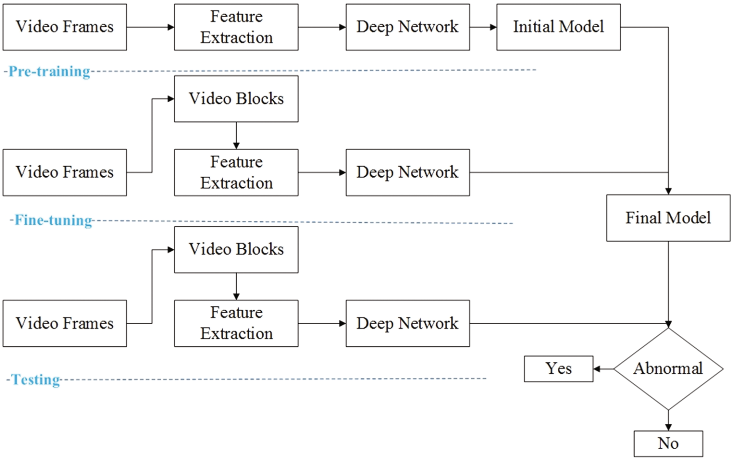

Fig. 1. Framework of the proposed method. (a) Pre-training: videos are irrelevant with abnormal behavior. (b) Fine-tuning: videos contain few abnormal samples and are decomposed into blocks. (c) Testing: videos are decomposed into blocks.

The selection of initial weights has great influence on the networks, furthermore, the detection of abnormal event is more depending on better weights. We pre-train our model on vast irrelative videos which are easy to obtain. Abnormal behavior often occurs in a small area while the entire scene covers large area in surveillance videos, we propose a patch strategy by decomposing video into blocks which avoids that local abnormal behavior weakened by global information. Video blocks taken as new inputs of networks to refine the weights. When performing abnormal event detection, the test videos will also be processed as before. Finally, the experiential rule is used to classify the video frames into an abnormal event or normal one.

3 The Proposed Method

3.1 Spatio-Temporal Deep Networks

The convolutional architecture was introduced in [16], since the early 2000s, CNN has been applied with great success to the detection, segmentation and recognition of objects in images. There are four key ideas behind CNN that take advantage of the properties of natural signals: local connections, shared weights, pooling and the use of many layers. CNN is a deep model which possesses strong ability of self learning.

The traditional CNN performs 2D convolution which only computes spatial information. In the application of video, time domain information is necessary. We use 3DCNN to obtain the time domain information, this kind of hierarchical learned high-level features are much stronger than the low-level temporal feature such as optical flow.

3DCNN applies 3D convolution instead of 2D convolution in CNN. 3D convolution extends the 2D convolution by which computes the convolution in temporal dimension. In 3DCNN, we apply it based on Eq. (1) and the details refer to [12],

$$v_{st}^{xyz} = f(b_{st} + \sum_{m} \sum_{p=0}^{P_s-1} \sum_{q=0}^{Q_s-1} \sum_{r=0}^{R_s-1} w_{stm}^{pqr} v_{(s-1)m}^{(x+p)(y+q)(z+r)}), \tag{1}$$

where, the value of an unit at position (x, y, z) in the t-th feature map in the s-th layer, denoted as v_{st}^{xyz}, b_{st} is the bias for this feature map, $w_{stm}^{pqr}$is the value at the position (p, q, r) of the kernel connected to the mth feature map.

Our spatio-temporal deep networks have five pathways and different low-level features are the input of each pathway. In addition, the five pathways have the same structure based on [12]. After then, the uniform feature representation is obtained by fully connected layer from five pathways. Ultimately, softmax is applied to classify the outputs. Figure 2 shows the framework of the networks. The structure of mixture multi-channels obtains variant different properties from original videos. Therefore, the extracted features contain more hidden information with better representation of video content.

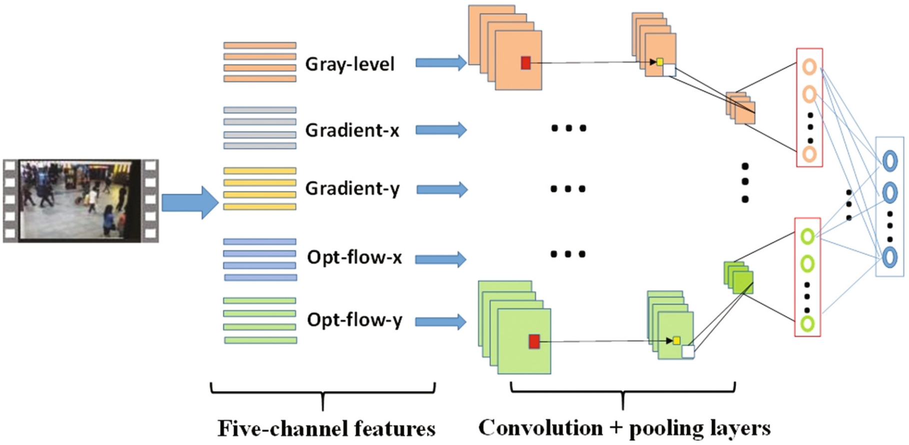

Fig. 2. The framework of networks.

Although CNN was designed for processing a huge number of training samples, we find that if the related class and unrelated class have some common properties, CNN can transfer the model pre-trained on the plenty of samples of unrelated class to the target class by fine-tuning on few related samples. In the detection of events, whether abnormal event or normal event, both of the motion representations are similar in videos. Therefore, we pre-train our deep networks with normal event videos and fine-tune on few abnormal events which demonstrated to be effective.

3.2 Patch Strategy

Considering abnormal behavior only locates in a certain area which accounts for a small proportion in surveillance video image, local event is likely to be weakened by global information if process the whole image directly. So as to solve this problem well, we develop a patch strategy. We use a sequence of N fixed-length clips without overlapping frames to characterize video, each clip contains m frames, the k-th frame in the C-th clip is divided into $d_1 \times d_2$ patches, where $i = 1, 2, \ldots, d_1$, $j = 1, 2, \ldots, d_2$, $k = 1, 2, \ldots, m$, $C = 1, 2, \ldots, N$. Then combining multiple patches together at same location in temporal dimension, the patch clip denoted as $P^n(i, j)$ and labeled simultaneously. The Dividing process as shown Fig. 3.

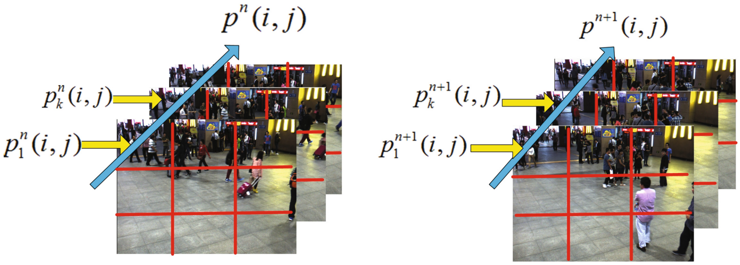

Fig. 3. Dividing process from continuous frames.

The testing videos are decomposed into blocks according to above method, then we apply the learned model for detecting abnormal events in videos. To prevent the false detection, we introduce an experiential rule works as follows, if k continuous clips belong to the category of abnormality simultaneously, it is regarded as abnormal event follows Eq. (2).

$$Result = \begin{cases} abnormal, p^n \bigwedge p^{n+1} \bigwedge \cdots \bigwedge P^{n+k-1} = 1 \\ normal, others \end{cases} \tag{2}$$

Where P^n denotes the result of detection at the same location, we denote $P^n = 1$ with respect to abnormal event. The value of k is set to 2 empirically.

4 Experimental Results

This section details our experimental protocols and describes the three videos datasets. We focus on the RICOH data to evaluate the developed method for abnormal detection. Meanwhile, we also compare with previous method which was evaluated in the same dataset [11]. Our method is implemented over CAFFE (Convolutional Architecture for Fast Feature Embedding) with respect to an open-source and high-efficiency deep learning framework.

4.1 Dataset

The Fudan-Columbia video Dataset (FCVID) consists of 91,223 Web videos annotated manually according to 239 categories. We choose six kinds of events in FCVID dataset to pre-training. RICOH dataset consists of two view points from different cameras (TYZX camera and Point-Gray camera) and includes amount of crowd events. The publicly available dataset from University of Minnesota (UMN) is used to detect escape event.

4.2 Experiments

In this experiment, ten frames of size 64 × 64 as inputs ($64 \times 64 \times 10$) to this model. The parameters of the nets are set as $C1(11 \times 11 \times 4) - S2(2 \times 2) - C3(7 \times 7 \times 3) - S4(3 \times 3) - C5(7 \times 7 \times 1) - FC6(256) - FC7(2)$. $C(H \times W \times T)$ represents convolutional layer, in this layer, H, W, T represent the height, width and size of temporal dimension in kernel respectively. After each convolutional processing, a *tanh* function is stacked as activation function. $S(H \times W)$ as the layer of subsampling, H and W represent the height and width of pooling. After the multiple layers of convolution and subsampling, we apply fully connected layers $FC(N)$ which consist of N feature maps. Obviously, the outputs, in the last layer, include two cases of abnormal and normal. The layers from $C1$ to $C5$ are common structures and $FC6$, $FC7$ are fusion layers. Finally, softmax shows promising capability in classification. All the parameters are initialized randomly and trained by back-propagation algorithm based on [10].

Figure 4 shows the results of crowd detection on RICOH dataset. In order to embody the effect of proposed method, we obtain the statistical result by decomposing the Point-Gray data into clips and each clip consists of 20 frames. The experiment shows that the accuracy of our method is 0.9658 and the F-measure is 0.9658. Both accuracy and F-measure are slightly better than method of [11], which are 0.9607 and 0.9578. It is worth noting that the model is trained and tested in different positions by two different cameras, which implies our method is robust for variation of viewpoints. However, the method in [7,21] is specially designed for specific scenes.

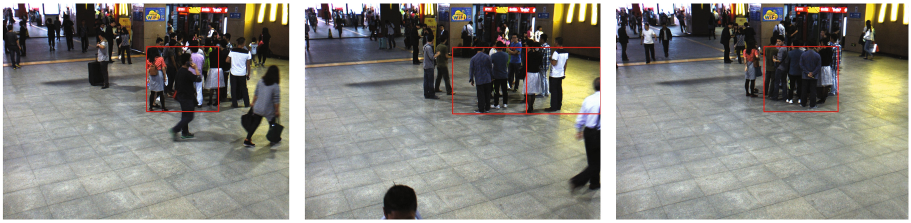

Fig. 4. The localization of crowd behaviors in the frames.

We also test the model over UMN dataset, Fig. 5 shows the result of escape event detection. The escape event can be detected accurately, but along with

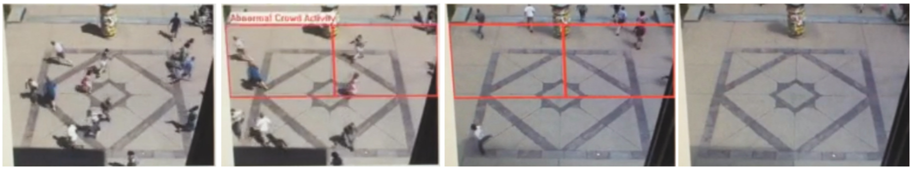

Fig. 5. The localization of escape behaviors in the frames.

delay. On the one hand, the experiential rule will lead to delay. On the other hand, the velocity is fast when human escape suddenly.

5 Conclusion

In this paper, we have developed a method for abnormal event detection based on deep learning. The proposed method pre-trains deep spatio-temporal networks over unrelated dataset and shows promising capability in abnormal event detection of small sample problem. We also introduce an experiential rule by which improving the effect of classification. Extensive experimental results illustrate that our method is robust of variant scenes. As part of future work, we plan to add statistic model which represents priori knowledge into deep networks, further improve its ability of small sample events detection.

References

1. Bhattacharya, S.: Recognition of complex events in open-source web-scale videos: a bottom up approach. In: Proceedings of the 21st ACM International Conference on Multimedia, pp. 1035–1038. ACM (2013)
2. Cho, S.H., Kang, H.B.: Abnormal behavior detection using hybrid agents in crowded scenes. Pattern Recogn. Lett. **44**, 64–70 (2014)
3. Cong, Y., Yuan, J., Liu, J.: Abnormal event detection in crowded scenes using sparse representation. Pattern Recogn. **46**(7), 1851–1864 (2013)
4. Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2625–2634 (2015)
5. Foggia, P., Saggese, A., Strisciuglio, N., Vento, M.: Exploiting the deep learning paradigm for recognizing human actions. In: 2014 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 93–98. IEEE (2014)
6. Gan, C., Wang, N., Yang, Y., Yeung, D.Y., Hauptmann, A.G.: Devnet: a deep event network for multimedia event detection and evidence recounting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2568–2577 (2015)
7. Gnanavel, V.K., Srinivasan, A.: Abnormal event detection in crowded video scenes. In: Satapathy, S.C., Biswal, B.N., Udgata, S.K., Mandal, J.K. (eds.) Proceedings of the 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2014. AISC, vol. 328, pp. 441–448. Springer, Cham (2015). doi:10.1007/978-3-319-12012-6_48

8. Graves, A.: Generating sequences with recurrent neural networks. arXiv preprint arXiv:1308.0850 (2013)
9. Gu, X., Cui, J., Zhu, Q.: Abnormal crowd behavior detection by using the particle entropy. Optik-Int. J. Light Electron Opt. **125**(14), 3428–3433 (2014)
10. Hansen, L.K., Salamon, P.: Neural network ensembles. IEEE Trans. Pattern Anal. Mach. Intell. **10**, 993–1001 (1990)
11. Hu, D., Meng, B., Fan, S., Cheng, H., Yang, L., Ji, Y.: Real-time understanding of abnormal crowd behavior on social robots. In: Ho, Y.-S., Sang, J., Ro, Y.M., Kim, J., Wu, F. (eds.) PCM 2015. LNCS, vol. 9315, pp. 554–563. Springer, Cham (2015). doi:10.1007/978-3-319-24078-7_56
12. Ji, S., Xu, W., Yang, M., Yu, K.: 3d convolutional neural networks for human action recognition. IEEE Trans. Pattern Anal. Mach. Intell. **35**(1), 221–231 (2013)
13. Jiang, Y.G., Bhattacharya, S., Chang, S.F., Shah, M.: High-level event recognition in unconstrained videos. Int. J. Multimedia Inf. Retriev. **2**(2), 73–101 (2013)
14. Kim, J., Grauman, K.: Observe locally, infer globally: a space-time MRF for detecting abnormal activities with incremental updates. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, pp. 2921–2928. IEEE (2009)
15. Le, Q.V., Zou, W.Y., Yeung, S.Y., Ng, A.Y.: Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3361–3368. IEEE (2011)
16. LeCun, Y., Kavukcuoglu, K., Farabet, C., et al.: Convolutional networks and applications in vision. In: ISCAS, pp. 253–256 (2010)
17. Mehran, R., Oyama, A., Shah, M.: Abnormal crowd behavior detection using social force model. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, pp. 935–942. IEEE (2009)
18. Popoola, O.P., Wang, K.: Video-based abnormal human behavior recognition - a review. IEEE Trans. Syst. Man Cybern. Part C: Appl. Rev. **42**(6), 865–878 (2012)
19. Qin, L., Ye, Y., Su, L., Huang, Q.: Abnormal event detection based on multi-scale markov random field. In: Zha, H., Chen, X., Wang, L., Miao, Q. (eds.) CCCV 2015. CCIS, vol. 546, pp. 376–386. Springer, Heidelberg (2015). doi:10.1007/978-3-662-48558-3_38
20. Sun, L., Jia, K., Yeung, D.Y., Shi, B.E.: Human action recognition using factorized spatio-temporal convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4597–4605 (2015)
21. Wang, T., Snoussi, H.: Detection of abnormal events via optical flow feature analysis. Sensors **15**(4), 7156–7171 (2015)
22. Yang, Y., Shah, M.: Complex events detection using data-driven concepts. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7574, pp. 722–735. Springer, Heidelberg (2012). doi:10.1007/978-3-642-33712-3_52
23. Zhang, Y., Qin, L., Yao, H., Huang, Q.: Abnormal crowd behavior detection based on social attribute-aware force model. In: 2012 19th IEEE International Conference on Image Processing (ICIP), pp. 2689–2692. IEEE (2012)
24. Zhou, S., Zhang, Z., Zeng, D., Shen, W.: Abnormal events detection in crowded scenes by trajectory cluster. In: International Symposium on Precision Engineering Measurement and Instrumentation, p. 944614. International Society for Optics and Photonics (2015)